Comprehensive genome and transcriptome analysis reveals genetic basis for gene fusions in cancer

Nuno A. Fonseca[1*], Yao He[2*], Liliana Greger[1], PCAWG3, Alvis Brazma[1], Zemin Zhang[2]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK;

[2]Peking-Tsinghua Centre for Life Sciences, BIOPIC, and Beijing Advanced Innovation Centre for Genomics, Peking University, Beijing, 100871, China

*Joint first authors

Online Methods

Data

We analysed RNA-seq data from 1188 donors with matched whole genome sequencing (WGS) data [1] derived from samples of 22 TCGA studies (755 tumour samples) and 7 others ICGC studies (454 tumour samples). In total RNA-seq from 1359 samples were analysed, encompassing 1209 samples from tumours and 150 from normal or adjacent tissues (Supplementary Fig. 1). Non-primary tumour samples were excluded from recurrency analysis (24 samples). In addition RNA-seq data from GTEx [2] (version phs000424.v4.p1) was used to complement the panel of normal tissues. GTEx RNA-seq data were realigned to the PCAWG reference using the PCAWG RNA-seq SOP [3]. The Human genome (GRCh37.p13) and annotation were obtained from Gencode (release 19). Cancer gene census list was compiled from the Catalogue of Somatic Mutations in Cancer version 80 [4].

The following data sets were obtained from the PCAWG repository in Synapse (https://www.synapse.org/): gene expression (Tophat2/Star FPKM) <syn5553985>; transcript expression <syn5974793>; clinical and histological data <syn7253568, syn7253569>; RNA-seq metadata <syn7416381>; consensus somatic structural variants (version 1.6 <syn7596712>) [5]; somatic driver mutations <syn9758012>; consensus somatic mutation calls (SNVs and Indels) <syn7118450>; and copy number alterations <syn7499507>; PCAWG's cancer driver genes (candidate release 24/4/2017) <syn9758012>. Further details about the data can be found elsewhere [1].

Gene fusion detection using RNA-Seq

For each aliquot with paired-end RNA-seq reads FusionCatcher [6] (version 0.99.6a) was applied to the raw reads, with the genome reference. The "-U True; -V True" runtime options were used. For each aliquot with single-end RNA-seq reads, STAR-Fusion [7] (version 0.8.0) was applied to the raw reads, with the same reference genome and gene models and with default settings.

In parallel FusionMap (version 2015-03-31) [7] was applied to all unaligned reads from the PCAWG aligned TopHat2 RNA-seq BAM files [3] for each aliquot to detect gene fusions with the following non-default options values: MinimalHit = 4; OutputFusionReads = True; RnaMode = True; FileFormat = BAM.

The output of both tools was post-processed so that the fusions detected and respective breakpoints and numbers of supporting reads were passed to the next step as depicted in Supplementary Fig. 2. The following filters were applied independently to the fusions detected by the two tools: i) the fusion has 4 or more breakpoint-spanning reads; ii) gene partners are not similar (paralogs or pseudogenes according to Ensembl Biomart); iii) the fusion was not detected in GTEx or in PCAWG's normal samples. This filter relied on analysis of GTEx and normal samples in the PCAWG cohort using the same fusion detection pipeline.

The two sets of fusion that passed the above criteria were then merged as follows. First, a fusion was retained if it was detected in a sample by both fusion detection pipelines and the breakpoint locations were consistent. Then two further constraints were applied: i) a fusion was detected by both fusion detection tools in at least one sample; or ii) a fusion detected by one of the methods had a matched SV in at least one sample. For integration with SVs, a fusion was considered to match a structural variant (SV) if the absolute distance between the fusion breakpoints and SV breakpoints did not exceed 500 KB (the distance was considered infinite when the chromosomes of the fusion and SV breakpoint differ). When there was no evidence for a direct SV fusion, the search was expanded to look for composite fusions. In this case an exhaustive search was performed to look for two SVs with breakpoints close to the fusion breakpoints and with an effective distance smaller than 250KB (Figure 3A). For instance, let F_1 and F_2 be the fusion breakpoints locations of a fusion F, and A_1, A_2, B_1, B_2 be the breakpoint locations of two structural variants (A and B), and dist(x,y) the distance between two breakpoint locations (infinite if the chromosomes are different). The search would try to find SVs A and B such that: a) $dist(F_1,A_1)$<=500000; b) $dist(F_2,B_1)$<=500000; c) $dist(A_2,B_2)$<=250000. Note that

there are other valid combinations implemented (e.g., where the $dist(F_1,A_2)$<=500000) but for conciseness only this example is here presented.

This resulted in detection of 3540 fusion events, from these 2268 were detected by both FusionCatcher/STAR-Fusion and FusionMap (from these 1821 had SV support) and 1112 were detected by only one method and had SV support. All fusions are available in Synapse (syn10003873).

Transcript quantification

Transcript quantification was estimated for each sample using Kallisto (v0.42.4) [9]. The set of protein coding transcripts found in Gencode 19 was used as the reference. For each gene and sample, a transcript was defined as *dominant* if its expression was reported at least 2-fold higher than the expression of the second most abundant transcript [10]. Expression of putative fusion transcripts was also estimated with Kallisto but with an extended set of transcripts: a non-redundant set of fusion transcripts together with the set of protein coding transcripts from Gencode 19.

Fusion reading frame prediction and domain annotation

For each gene fusion candidate, the entire chimeric transcript sequence was assembled. First, we selected the dominant isoform for the fusion gene as the reference transcript; if a gene in an aliquot did not have a dominant isoform, the longest CDS transcript was selected. Second, based on fusion gene breakpoint coordinates, chimeric sequences based on the isoforms were selected and splicing junctions of each gene were reconstructed. Finally, based on the positions of the junctions on the transcripts we tagged them as coding (CDS) if the junctions fell within the coding boundaries and as 5' un-translated region (5UTR) or 3' un-translated region (3UTR) if they fell outside the coding boundaries. The positions of breakpoints at the codon level were related to the lengths of the open reading frames. Based on the predicted reading frame potential we classified fusions into several groups: *in-frame, out-of-frame, 5'UTR->CDS, CDS->5'UTR, 5'UTR->5'UTR, 3'UTR->CDS, CDS->3'UTR, 5'UTR->3'UTR, 3'UTR->5'UTR, 3'UTR->3'UTR*. For fusions involving non-coding genes the reading frames were assigned to the category *Other*s. The structural domains for each predicted fusion protein were annotated using the Uniprot and Pfam domain tracks downloaded from the UCSC table browser. For a number of recurrent fusions such as *DCAF6-MPZL1* and *GNS-NUP107*, manual literature search was used to refine the the definition of protein domains for the leucine-zipper motif.

Identification of sequence features and DNA repair mechanisms associated with bridged fusions

SV densities and mutation densities were calculated by sampling each 1 Mbp chromosomal window across all samples containing bridged fusions. The chromothripsis calls for each sample are described in [11] . Consensus somatic mutation calls were used to characterize kataegis for each sample as follows. A genomic window would be considered to have kataegis if it had at least 6 consecutive mutations with an average inter-mutation distance of less than 10 kb. We used the MEME suite (http://meme-suite.org/) to identify de novo motifs that might be enriched in the vicinity of pillar breakpoints. Specifically, we used the DREME algorithm with an E-value threshold of 0.05 to search for motifs in the 50 bp window of pillar breakpoints [12]. Controls sequences were based on the randomly shuffled original input sequences. Repeat elements that were used to overlap with fusion breakpoints were obtained from the UCSC RepeatMasker track, and common fragile sites were downloaded from [13] and were converted to the hg19 genome coordinates by liftover.

The DNA repair-related mechanisms were predicted based on the homology and mutation features at both sides of the pillar breakpoints for each bridged fusion. Microhomology sequences for each breakpoint were extracted from the consensus somatic structural variants. The classification criteria was defined previously [14, 15, 16]. The mutation rate in the vicinity of the 100kb window of the bridged regions were calculated using circlize R package [17] to infer hypermutation status using consensus somatic mutation calls.

Previously reported gene fusions

We used ChimerDB 3.0 [18] as a reference of previously reported gene fusions. It contains 32,949 fusion genes splitted into three groups:
- ***KB***: 1,067 fusion genes manually curated based on public resources of fusion genes with experimental evidences;
- ***Pub***: 2,770 fusion genes obtained from text mining of PubMed abstracts.
- ***Seq***: archive with 30,001 fusion gene candidates from deep sequencing data. This set includes fusions found by re-analysing the RNA-seq data of the TCGA project encompassing 4,569 patients from 23 cancer types .

Throughout the manuscript we refer previously reported fusions as "high confidence" if they are part of the KB and Pub groups.

Statistical analysis

All statistical analyses were performed using R 3.3. The association of the genes with different pathways was performed using the Pathview package [19] version 1.14.0 analysis, the background included all genes involved in gene fusions. The STRING database [20] (version 10) and the STRINGdb R package (version 1.14.0) was used as the reference of known protein-protein interactions and the tool for visualizing protein-protein interactions and to perform promiscuous gene fusion partner overlap enrichment.

References

1. PanCancer Analysis of Whole Genomes . (in preparation)
2. GTEx Consortium. "The Genotype-Tissue Expression (GTEx) pilot analysis : Multitissue gene regulation in humans." Science 348.6235 (2015): 648-660.
3. PCAWG-3. Pan-Cancer Study of Recurrent and Heterogeneous RNA Aberrations and Association with Whole-Genome Variants. (in preparation).
4. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43,** D805-811 (2015).
5. Patterns and mechanisms of structural variation in human cancers. (in preparation)
6. D. Nicorici, M. Satalan, H. Edgren, S. Kangaspeska, A. Murumagi, O. Kallioniemi, S. Virtanen, O. Kilkku, FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data, bioRxiv, Nov. 2014, DOI:10.1101/011650
7. Haas, B. *et al.* STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv* 120295 (2017). doi:10.1101/120295
8. Ge, Huanying, et al. "FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution." Bioinformatics 27.14 (2011): 1922-1928.
9. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34,** 525–527 (2016).
10. Gonzàlez-Porta, Mar, et al. "Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene." *Genome biology* 14.7 (2013): R70.
11. PCAWG-6 Comprehensive characterization of chromothripsis across 2600 human cancers using whole-genome sequencing. Isidro Cortés-Ciriano, Ruibin Xi, June-Koo Lee, Dhawal Jain, Youngsook L. Jung, Dmitry Gordenin, Peter J. Park* and PCAWG-6. (in preparation)

12. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27,** 1653–1659 (2011).
13. Le Tallec, Benoît, et al. "Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes." *Cell reports* 4.3 (2013): 420-428.
14. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17,** 224–238 (2016).
15. Ottaviani, D., LeCain, M. & Sheer, D. The role of microhomology in genomic structural variation. *Trends Genet.* **30,** 85–94 (2014).
16. Sinha, S. *et al.* Microhomology-mediated end joining induces hypermutagenesis at breakpoint junctions. *PLoS Genet.* **13,** e1006714 (2017).
17. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize Implements and enhances circular visualization in R. *Bioinformatics* **30,** 2811–2812 (2014).
18. Lee, Myunggyo, et al. "ChimerDB 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining." *Nucleic acids research* 45.D1 (2017): D784-D789.
19. Luo, W. and Brouwer C., Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics, 2013, 29(14): 1830-1831.
20. Szklarczyk, Damian, et al. "STRING v10: protein–protein interaction networks, integrated over the tree of life." *Nucleic acids research* (2014): gku1003.